Advanced Algorithms in Bioinformatics (P4)

Sequence and Structure Analysis

Freie Universität Berlin, Institut für Informatik
Prof. Dr. Knut Reinert, Sandro Andreotti
Sommersemester 2009

1. Review, 2009-05-29

Name, Vorname	Matrikelnummer

Zur Bearbeitung des Reviews stehen Ihnen 50 Minuten zur Verfügung. Jeder Punkt entspricht in etwa einer Minute.

Geben Sie auf diesem Titelblatt und auf allen eventuell zusätzlich abgegebenen Blättern ihren Namen und ihre Immatrikulationsnummer an.

Schreiben Sie ihre Lösungen direkt auf die entsprechenden Aufgabenbögen. Sollte dort der Platz nicht ausreichen und Sie weitere Blätter benötigen, vermerken Sie dies bitte, damit wir auch den Rest ihrer Antwort finden und bei der Bewertung berücksichtigen können. Am Ende des Reviews sind sämtliche Aufgabenblätter wieder abzugeben.

Ergebnis:

Aufgabe	maximal	erreicht
1	12	
2	12	
3	11	
4	15	
$\sum$	50	

Exercise 1. 4+2+6=12 Punkte

1. Explain the idea of the Horspool algorithm for exact pattern matching.
2. How does the performance depend on the alphabet size? Explain why?
3. Apply the Wu-Manber algorithm to search for the three patterns `AGGTG`, `ACGAT`, and `TATAC` in the text `AGACACGCTCTATAC`. Use a Block size of two and the following hash function H:

   $$H(\texttt{AA}) = H(\texttt{AC}) = 0, H(\texttt{AG}) = H(\texttt{AT}) = 1, H(\texttt{CA}) = H(\texttt{CC}) = 2, \ldots H(\texttt{TG}) = H(\texttt{TT}) = 7$$

   for both tables.

Exercise 2. 6 + 4 + 2 = 12 Punkte

The Myers Bitvector algorithm uses binary encoding of the dynamic programming matrix.

1. Use the bitvectors to fill out the dynamic programming matrix

$$
\begin{array}{lcl}
VN_1 & = & 000000 \\
VP_1 & = & 111110 \\
D0_2 & = & 111110 \\
HN_3 & = & 111100 \\
HP_3 & = & 000010
\end{array}
$$

		t_1	t_2	t_3
	0	0	0	0
p_1	1			
p_2	2			
p_3	3			
p_4	4			
p_5	5			
p_6	6			

2. Below you find the pseudocode of the Myers Bitvector algorithm. Complete the lines 11 and 13

3. How can you modify the algorithm to compute edit distance (global alignment) instead of the semi-global alignment?

```
1  // Searching
2  for pos ∈ 1...n do
3      X = B[t_pos] | VN;
4      D0 = ((VP + (X & VP)) ∧ VP) | X;
5      HN = VP & D0;
6      HP = VN | ∼(VP | D0);
7      X = HP ≪ 1;
8      VN = X & D0;
9      VP = (HN ≪ 1) | ∼(X | D0);
10     // Scoring and output
11     if ..... ≠ 0^m
12        then score += 1;
13         else if ..... ≠ 0^m
14                 then score −= 1;
15              fi
16     fi
17     if score ≤ k report occurrence at pos fi;
18 od
```

Exercise 3. 3+2+4+2=11 Punkte

1. State the q-gram Lemma and prove it.
2. State the q-gram Lemma for gapped shapes.
3. Is it tight for gapped shapes? Prove your answer by example. Use a gapped shape with $|Q| \geq 3$.
4. What are the advantages and disadvantages of gapped shapes compared to contiguous shapes?

Exercise 4. 4+5+6=15 Punkte

1. The general idea of the Manber Myers Algorithm is prefix doubling. Explain how this approach allows for sorting the suffixes without doing character comparisons (except in the first stage).

2. The Manber Myers algorithm uses five arrays for the construction of the suffix array for a text T in time $O(|T| \log |T|)$. Explain the role of the arrays suftab, sufinv, count, bh, and b2h during the construction algorithm.

3. Assume your suffix array is sorted with respect to the first two characters. Below you find some entries of the sufinv array. Fill out the blank entries. Everything you need to know is that $S_6 = abcab\$$ and $\Sigma = \{a, b, c\}$ (Hint: Use the available entries to reconstruct the complete sequence.)

i	sufinv[i]	Bh[i]
0	3	1
1		1
2	4	0
3		1
4	7	1
5		0
6	1	1
7	5	1
8	8	0
9		1
10		1
11		1